View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repository@Nottingham

Effects of Display Design on Signal Detection in Flash Flood Forecasting

Elizabeth M. Argyle [a, b, c *], Jonathan J. Gourley [b], Chen Ling [d],
Randa L. Shehab [a], and Ziho Kang [a]

[a] School of Industrial and Systems Engineering
University of Oklahoma
202 W. Boyd St., Room 124
Norman, OK 73079 USA

[b] NOAA/National Severe Storms Laboratory
120 David L. Boren Blvd.
Norman, OK 73072 USA

[c] Cooperative Institute for Mesoscale Meteorological Studies
University of Oklahoma
120 David L. Boren Blvd., Suite 2100
Norman, OK 73072 USA

[d] Department of Mechanical Engineering
The University of Akron
Akron, OH 44325 USA

[Accepted version – the published version is available through the International Journal of Human-Computer Studies, March 2017, 99, p. 48-56 at https://doi.org/10.1016/j.ijhcs.2016.11.004]

*Corresponding author: Elizabeth M. Argyle, +44 (0)115 95 86439,
elizabeth.argyle@nottingham.ac.uk

Abstract

The Flooded Locations and Simulated Hydrographs (FLASH) project is a suite of tools that use weather radar-based rainfall estimates to force hydrologic models to predict flash floods in real-time. However, early evaluation of FLASH tools in a series of simulated forecasting operations, it was believed that the data aggregation and visualization methods might have contributed to forecasting a large number of false alarms. The present study addresses the question of how two alternative data aggregation and visualization methods affect signal detection of flash floods. A sample of 30 participants viewed a series of stimuli created from FLASH images and were asked to judge whether or not they predicted significant or insignificant amounts of flash flooding. Analyses revealed that choice of aggregation method did affect probability of detection. Additional visual indicators such as geographic scale of the stimuli and threat level affected the odds of interpreting the model predictions correctly as well as congruence in responses between national and local scale model outputs.

Keywords: data aggregation; visualization; weather forecasting; flash flooding; human factors; decision making; signal detection; situation awareness

1. Introduction

In the field of weather forecasting, computational modelers are under pressure to provide actionable information to end users at increasingly local levels, pushing gridded forecasting systems to hyper-resolution scales (Wood et al., 2011; Beven et al., 2015). Although the capability to predict weather phenomena at small scales continues to develop, operational technology often limits display capacity. Large high-resolution displays have been shown to overcome data abstraction limits while enabling users to engage in exploratory data analysis (Lehmann et al., 2011). However, current operational forecasting display systems are frequently based on the multi-screen desktop setup, and meteorological visualization environments are constrained to comparatively low resolution displays.

1.1 The Flooded Locations and Simulated Hydrographs (FLASH) Project

One such set of gridded forecasting products is the Flooded Locations and Simulated Hydrographs (FLASH) project. FLASH is a suite of real-time tools that use weather radar-based rainfall estimates to force hydrologic models to predict flash floods. The tools provide environmental information related to flash flood risk to professional forecasters, and the simulation models are designed to overcome several limitations of existing prediction systems (Gourley et al., 2016). The grid underlying each FLASH product covers a spatial extent of the continental United States at a horizontal resolution of 1 km. The hydrologic model calculates a return period, a measure of flash flood risk, for every cell within the grid. In hydrologic terms, a return period is the average length of time for a certain threshold of flooding to be reached (Mays, 2010). Potential FLASH users include forecasters at both the national and regional scales in the United States, including, but not limited to, National Weather Service Weather Forecast Offices (WFOs), River Forecast Centers (RFCs), and national centers. The tools are intended to assist forecasters to identify areas of dynamic flood risk across the country and, in turn, to predict specific threats.

When this work took place in 2013, the FLASH product suite was in development and experimental simulations were publicly displayed through a website. The website's visualization template was originally developed to display interactive data related to the National Mosaic and Multi-Sensor Quantitative Precipitation Estimates (NMQ) system (Zhang et al., 2011). When applied to the FLASH return period visualization, the pre-existing algorithm aggregated grid cells as the user zoomed in and out. At the finest scale, all grid cells were visible, but as a user zoomed out to the national map, an overview of the data presented aggregated sets of grid cells within each pixel. However, a design challenge emerged at this stage: when showing the map of the entire continental United States, the website platform and some desktop-based display systems were not able to display each individual grid cell.

The original website displayed an overview of multiple grid cells with an aggregation algorithm to sample the maximum value out of a collection of at least 112 grid cells contained within one pixel. Predictions were displayed without any form of filtering first. In practice, while the true predicted return period values were presented when a viewer zooms in to a local level, the national view displayed an aggregated overview of the data by displaying the maximum value. An example of this phenomenon is shown in Figure 1. At the national level, this resulted in an occlusion effect, where lower return period values were occluded by the maximum values.

1.2 Motivation

In July 2013, the Hydrometeorological Testbed at the Weather Prediction Center (HMT-WPC) hosted the first Flash Flooding and Intense Rainfall (FFaIR) experiment (Barthold et al., 2015). The purpose of the experiment was to evaluate the utility of several experimental forecast models, including FLASH, with professional forecasters and weather researchers. During the testbed, forecasters predicted heavy rainfall and flash flooding using the operational and experimental computational model outputs. Throughout these activities, the researchers observed that the information visualization affected the forecasters' ability to interpret the FLASH data. Forecasters commented that their flash flood predictions turned into false alarms more frequently

in the experiment than during typical operations, which they attributed to FLASH's data aggregation algorithm. Based on these subjective comments, the researchers hypothesized that changing the aggregation algorithm would affect the rate of false alarm forecasts. In order to test this, the researchers created an alternative aggregation method which took the mean value of the grid cell predictions for a given subset (hereafter referred to as the "average-based aggregation algorithm").

The present study identified differences in terms of error rates when comparing maximum-based and average-based aggregation algorithms on the national-scale maps. This work expands upon a preliminary error rate analysis presented by Argyle et al. (2015). In addition, an analysis of response congruence was undertaken in order to determine the effects of the display condition on response accuracy across both levels of geographic scale (the national level and the zoomed-in, local level). From a design perspective, congruent decisions between levels of geographic scale are highly desirable. In FLASH, the national overview provides insight into environmental threats across the country to direct a forecaster's attention to at-risk regions. Likewise, a forecaster working at a local level may wish to examine a broader geographic region to determine potential future threats and broader environmental conditions. As such, congruent judgments between levels indicate the degree of fidelity between the abstracted overview and the individual grid cell predictions.

2. Related Work

Visualization design can have a great influence on decision making and performance in weather forecasting, which largely consists of detection and identification processes (Bowden et al., 2015). Detection and identification occur rapidly and are governed by cognitive structures such as long term memory, working memory, schema, mental models, attention, feature identification, and monitoring, among others (Adams et al., 1995; Endsley, 1995, 2015; Hoffman, 2015; Wickens, 2015). In addition to these factors, success in weather forecasting has been attributed to the forecaster's ability to acquire and maintain situation awareness (Quoetone et al.,

2001). As defined by Endsley (1995), situation awareness (SA) is the ability to perceive elements within a system, comprehend their significance, and project their meaning into the future in order to make a decision. Underlying the SA construct are personal factors and cognitive mechanisms, including visual information processing, cue detection, working memory, goals, preconceptions, background knowledge, and system design (Adams et al., 1995; Endsley, 1995, 2015; Hoffman, 2015).

In practice, detection is a function of factors including top-down processes, expectations, and background knowledge, aligning with Level 1 of Endsley's (1995) Model of SA, perception. Identification involves detecting an item and evaluating its fit into a categorical grouping, and it is also affected by experience and top-down processes (Endsley, 1995; Wickens and Carswell, 1997). Identification can be mapped to Level 2 of Endsley's (1995) Model of SA, or comprehension. While the third level of Endsley's 1995 Model of SA, projection, was determined to be outside the scope of the present study, future work could extend the present study's method from a detection and identification task to a projection task in which participants would have to choose whether or not a flash flood warning would be appropriate.

Detection and identification tasks can also be framed within the family of cognitive integration processes. Graph comprehension studies distinguish specific information extraction processes from information integration. In the former, a user has a goal to search and find a specific attribute in a visualization; in the latter process, a user may combine multiple attributes from a visualization in order to comprehend broader meanings and trends in the data (Ratwani et al., 2008). Due to the map-based format of many data sources used in weather forecasting, information integration is a fundamental activity for a forecaster to be able to develop SA. In example, examining a FLASH return period value assigned to a single grid cell provides much less meaning than evaluating the overall trends and gradients over broader geographic scales. Information integration in graph comprehension has been viewed as iterative processes of pattern recognition and interpretation, in which features are detected and, ideally, understood; as graph

complexity increases, more iterations of the integration process are required (Ratwani et al., 2008).

In addition to situational and cognitive factors, the warning decision process is affected by factors such as forecaster experience and task-relevant knowledge, risk tolerance, perceptions and beliefs about environmental states, confidence, software issues, and spatial ability (Heinselman et al., 2012; Smallman and Hegarty, 2007). Likewise, information format can impact forecaster performance. Studying the effects of variable update frequency from phased-array radar data, Bowden et al. (2015) observed that the probability of detection for detecting and identifying a severe hail or wind threat increased as update frequency increased. Anchoring, or cognitive bias based on initial information received prior to making a decision, may also play a significant role in threat assessment in meteorological decision making. Joslyn et al. (2011) observed that by providing upper bounds of uncertainty to forecast data, decision makers predicted a higher value than when provided with a lower estimate. In relation to aggregated overviews, this suggests that a relationship may exist among anchoring, threat detection, and choice of aggregation algorithm.

One challenge in designing geospatial visualizations is that of reducing selection occlusion, particularly in aggregated overviews of data (Shrestha et al., 2014). Overviews permit users to view the data in order to see the big picture, often as a summary view or as a zoomed out view of the dataset. Zooming functions allow users to focus on particular points or subsets of the data in order to identify specific information attributes (Shneiderman, 1996). Visualization systems have varied in their approaches to providing aggregated overview of spatial data, and applications have included semantic zooming through an aggregated information space (Bederson et al., 1996), spatial visualizations of movement (Andrienko and Andrienko, 2011), hierarchical system representations (Elmqvist and Fekete, 2010; Lehmann et al., 2011), and power generation systems visualizations (Overbye and Weber, 2000).

In addition to conveying patterns in the data to users, overviews may also have utility during search-based activities. Information foraging theory proposes that humans seek information that is both salient and suited to their goals (Pirolli and Card, 1999). Using cognitive task analysis to assess the sensemaking process, Pirolli and Card (2005) found that expert intelligence analysts engaged in cycles of foraging, sensemaking, and reality/policy assessment. In the foraging loop, analysts gathered information sources and tried to make inferences from the data. The theory describes such information as having "scent" that directs the decision maker's attention towards particular aspects of the data during the foraging loop (Pirolli and Card, 1999). In an extension of IFT to foraging on websites, Fu and Pirolli (2007) confirmed that the scent-based foraging model described user interactions with the web-based information sources. Aggregation in overviews may provide scent to direct a user's attention towards salient points in the dataset, directing users to areas of the display that merit deeper inspection.

In order to develop overviews that adequately represent their underlying dataset, Elmqvist and Fekete (2010) recommend following the principles of visual summary and fidelity. The principle of visual summary states that the visual properties of the data aggregate should be representative of the individual data point members. However, certain aggregation methods can lead to loss of fidelity and misinterpretations of the visualization. Inadequacies of aggregation methods based on averages relate to the user's loss of knowledge about the variance within the aggregate (Elmqvist and Fekete, 2010). It is possible that interactive overviews may help to overcome fidelity loss. Additional design challenges for weather forecasting displays include improving visual discriminability (Dobson, 1979; Wickens and Carswell, 1997), highlighting meaningful information clusters to facilitate integration (Ratwani et al., 2008), and structuring the information landscape in a way that assists the user to achieve their goals in a hierarchical needs-based order (Hoffman and Woods, 2005; Trafton and Hoffman, 2007).

3. Method

3.1 Hypotheses

Forecaster comments from the 2013 FFaIR experiment led the researchers to hypothesize that an algorithm displaying the average of sampled grid cells (henceforth called the average-based display) would produce different task performance than the maximum-based display. In the Signal Detection Theory analysis, it was hypothesized that the maximum-based display would increase false alarm rates (i.e. the event was forecast but not observed) and hit rates (i.e. the event was correctly forecast and observed) over the average-based display. Due to the inherent priority given to visualizing the higher end of the color scale as well as the larger areal coverage of represented regions, it was thought that the maximum-based display would draw attention to severe events more rapidly than the average-based display would.

Furthermore, the display algorithms were compared against the principles of visual summary and fidelity. Here, it was assumed that a congruent judgment of an overview and its corresponding zoomed-in view of individual grid cells indicated high fidelity between the two views. In light of this, it was hypothesized that the average-based display would be associated with lesser congruence between views due to its flattening of the underlying variability.

3.2 Experimental Design

As a between-subjects independent variable, the aggregation algorithm differed across two levels: participants viewed visualizations created with either the maximum-based aggregation method or the average-based method. During the study, the stimuli were also presented on two levels: as a national overview on a map of the continental United States, and as a more finely-detailed regional map, zoomed in to an area of interest. However, it is important to note that while the national overviews used by the two participant groups varied by aggregation algorithm, the local images that participants viewed were identical between groups. The purpose of viewing identical local images was to identify whether any bias occurred in detection based on which level of national image a participant viewed first. This task also served to assess congruence in decision making between national and local views.

3.3 Task Materials

A series of forty image sequences was created by taking screen captures of the FLASH visualization. Each sequence consisted of one image of a FLASH model output visualized at a national overview level, and a second image of the same date and time, spatially focused to display a local, county- or state-level view of the flood event. Within each national overview stimulus, a white selection box indicated the region of interest. In each stimuli, the FLASH model plotted a return period value, measured in years and coded into a numerical color scale, for every grid cell in the map of the United States. The scale ranged from 0 to 200 years and plotted predicted values against four primary colors (gradients of green, yellow, red, and purple, in ascending order) and two secondary colors (black for 0-value predictions and grey for cells with missing data). An example of two image sequences, one in the average-based aggregation condition and the other in the maximum-based aggregation condition, is shown in Figure 2.

The stimuli were selected based on flash flooding events that occurred between April and July 2013 and were reported in the National Climatic Data Center Storm Events Database (National Climatic Data Center, 2014). When selecting the events from the database, the researcher categorized events into "severe" and "not severe" flash flooding. Unlike the Fujita scale, which estimates tornado scale, there is not yet a standardized metric for flash flooding severity. The research team used financial damages as a proxy for flood impact and arbitrarily defined severe flash flooding to be those that caused $500,000 or more of property and crop damage ($n = 20$, $\mu = \$10.38M$; $\sigma = \$22.82M$). Events that were placed in the "not severe" category had less than $500,000 of property and crop damage ($n = 20$, $\mu = \$38.75K$; $\sigma = \$84.59K$). During the tasks, geographic background knowledge was of some benefit to participants, although there was not necessarily a correlation between population density and high property damage. In future work, alternative proxy variables for flood impact should be examined.

In addition to the stimuli, the task materials included a training document which explained FLASH return period interpretation. The document contained examples of the visualization and described the modeling framework.

3.4 Procedure

Initially, participants received instruction about the study's purpose and tasks. After completing the informed consent process, participants read an excerpt from the FLASH training manual that explained how to interpret the FLASH visualization. This training document used pictorial examples to demonstrate appropriate interpretation of the FLASH return period color scale mapping. Participants were given the opportunity to ask questions about FLASH, how to interpret the display, and what the study would involve.

Once they felt comfortable with the FLASH interface, participants answered several demographic questions (age, gender, and academic classification). Following this, participants viewed the forty image sequences in a randomized order. In each sequence, the first image showed an event in FLASH on the national overview. The goal was to detect a signal (a high property damage threat) from the noise (a low damage threat). The distinction was based on the predicted return period values and the corresponding colors indicated in the scale. Participants were asked, "Based on the information that is modeled in this image, would you expect for this event to produce flash flooding with severe levels of property damage? (>$500,000)." Participants were instructed to base their judgments on their meteorological experience, any geographic knowledge they may have had, and their expectations based on predicted return period values. With this in mind, participants reviewed the image, and then pressed "y" for yes or "n" for no after making their decision. The following image was always a representation of the same weather event, but visualized at the local scale. The participants answered the same question about severity based on the new presentation. When participants finished with the final pair, they were debriefed.

3.5 Equipment

Images were randomly presented to participants using PsychoPy, an open-source software that allows researchers to present stimuli and collect response data from participants (Peirce, 2007). Each evaluation was conducted on an Asus A53U laptop with a 15-inch screen; each image was displayed at a size of 869x680 pixels. While professional forecasters often work with desktop systems, the laptop was able to present the stimuli at a similar resolution.

3.6 Participants

Thirty participants (19 male, 11 female) took part in the study. Participants were between the ages of 21-41 years old, with a mean age of 25.0 years and a median of 23 years. As the FLASH system was in development at the time of the study and thus not widely available to non-research personnel, participants had little to no experience using the FLASH return period visualization. However, in order to ensure that participants had relevant experience with interpreting environmental visualizations, participants were required to have more than one year of experience in meteorology. Participants were recruited from the student and post-doctoral population at the University of Oklahoma. Eligible individuals were either seeking a degree in meteorology or already possessed one.

3.7 Measures

Using a Signal Detection Theory framework, error rates were calculated from the response data from the detection task (McNicol, 2005). In traditional explanations of error rate analysis in weather forecasting, signal detection metrics are based on comparisons between the predictions and the actual outcomes. For example, a hit would occur when a flash flood was forecast and then actually occurred. A false alarm refers to an event in which a flash flood was forecast but then did not occur. Translated into the present study's framework, in which all stimuli visualized confirmed flash floods associated with reports, the explanation of error rates is instead based on correct identification of property damage level for NWS-verified flash floods.

The measure associated with the congruence analysis was frequency based. Judgment congruence occurred when a participant's national level stimuli judgment aligned with the

associated local level stimuli judgment. Decision making was examined in terms of judgments that were congruent and correct, congruent but incorrect, and incongruent.

4. Results

4.1 Error Rates, Sensitivity, and Bias

After collecting the participants' responses, the error rates in terms of the Signal Detection Theory framework were calculated for the severity judgment associated with the average-based and maximum-based display styles and for the national and local images. A sensitivity and bias analysis reflected that aggregation method influenced errors in the detection task. The sensitivity index (d') scores for the average-based display and maximum-based display were 0.88 (SD = 0.35) and 1.00 (SD = 0.35), respectively, $t(27.99) = 0.91$, $p = 0.37$. This indicated that there was no detectable difference between the discriminability of a severe flood signal between the two display types. A significant difference was also found in the biases associated with the two display algorithms, $t(15.53)$, $p < 0.01$: for the maximum-based display algorithm, a liberal bias of -0.74 was found (M = -0.74, SD = 0.97), and a conservative bias of 0.24 was found for the average-based display algorithm (M = 0.24, SD = 0.23). This can be interpreted to mean that participants in the maximum-based display condition were more likely to conclude that any stimulus contained a significant flood, while the participants in the average-based display condition were more likely to say a stimulus did not.

In further support of the sensitivity analysis, the error rate data were compared using t-tests, which showed a significant difference between the display methods. Participants using the maximum display produced a higher hit rate (M = 0.81, SD = 0.13) than those using average display (M = 0.57, SD = 0.12), $t(27.75) = 5.38$, $p < 0.001$. However, the average display minimized the false alarm rate (M = 0.25, SD = 0.08) when compared to the maximum display (M = 0.50, SD = 0.19), $t(18.95) = 4.63$, $p < 0.001$. A summary of the results is shown in Table 1.

Table 1. Error rate comparison between display types

	Hit Rate	False Alarm Rate
Average-based	0.57	0.25
Maximum-based	0.81	0.50
p-value	<0.001	<0.001

4.2 Effect of Display Design on Congruent Decisions

Congruent decisions were deemed either *congruent-correct* (a "yes/yes" response to an image sequence that represented a high damage level flood or a "no/no" response to an image sequence that represented an low damage level flood), *congruent-incorrect* (a "no/no" response to an image sequence that represented a high impact level flood or a "yes/yes" response to an image sequence that represented an low impact level flood), or *incongruent* (a "yes/no" or "no/yes" response, which by definition was always partially correct). Counts of congruent and incongruent decisions by display condition are shown in Tables 2 and 3.

A Chi-squared test of decision counts against display condition revealed a significant difference between the maximum-based aggregation algorithm and the average-based algorithm for judgment congruence. However, these differences were observed when assessing judgment congruence in relation to threat level. When judging images representing low property damage events, participants in the average display condition produced more congruent judgments (hits on both images within a given image sequence) than participants in the maximum display condition, χ^2 (2, N = 600) = 31.16, $p < 0.0001$. Conversely, when judging an image representing a high property damage event, participants in the maximum-based display condition produced more congruent hits and fewer congruent misses than those in the average-based display condition, χ^2 (2, N = 585) = 36.15, $p < 0.0001$.

Table 2. Counts of congruent and incongruent decisions by display condition for high-level property damage events

Threat Level: High Property Damage Events				
	Hit/Hit	Miss/Miss	Hit/Miss or Miss/Hit	Row Totals
Average-based	99 (33.0%)	84 (28.0%)	117 (39.0%)	300 (100.0%)
Maximum-based	148 (52.0%)	40 (14.0%)	97 (34.0%)	285 (100.0%)
p-value		< 0.0001		

Table 3. Counts of congruent and incongruent decisions by display condition for low-level property damage events

Threat Level: Low Property Damage Events				
	Hit/Hit	Miss/Miss	Hit/Miss or Miss/Hit	Row Totals
Average-based	204 (68.0%)	48 (16.0%)	48 (16.0%)	300 (100.0%)
Maximum-based	139 (46.3%)	66 (22.0%)	95 (31.7%)	300 (100.0%)
p-value		< 0.0001		

4.3 Did visual attributes affect the likelihood of a correct response?

The previous findings suggest that display condition and threat level did impact decision accuracy under certain conditions. Following the signal detection analysis, we hypothesized that an additional factor, visual distraction, may have affected participant judgments. Prior work has suggested that task-irrelevant features on geospatial displays may negatively impact task performance (Hegarty et al., 2012). Although participants were instructed to judge only the area within the white selection box on each stimulus, many of the stimuli contained visually distracting imagery of flood predictions outside the box. Each stimulus received a code to designate the amount of visual distraction as determined by the areal coverage of the grid cells

containing non-zero return period values. The new explanatory variable, areal size (s), was created with two levels: *small* and *large*. An example of a small-scale stimulus and a large-scale stimulus are shown in Figure 3 and Figure 4, respectively.

Due to the binary nature of participant responses to the threat detection task (0 = incorrect threat identification, 1 = correct threat identification), a logistic regression was chosen as the appropriate method to determine the relationship between display type, threat level, and areal size. Tests of the resulting model revealed that the full model containing all interactions were significant, suggesting that areal size, threat level, and display type influenced the likelihood of correctly identifying a flash flood threat, χ^2 (7, N = 1185) = 190.80, $p < 0.001$. This finding supports the earlier decision bias finding and provides further evidence that visual aspects of the stimuli affected the likelihood of correctly interpreting a threat. The odds ratio for the size variable is 0.011 (with a 95% confidence interval of 0.003 – 0.032); this is interpreted to mean that the odds of a participant giving a correct response to a stimulus that contained a large-scale event were 0.0110 times the odds of giving a correct response when viewing a stimulus that contained a small-scale event. In reverse, the odds of a participant correctly judging a small-scale event were approximately 90.909 times the odds of correctly judging a large-scale image. Likewise, the odds of a participant correctly judging a stimulus image that contained a significant threat of property damage were 0.145 times the odds of correctly judging an image with insignificant levels of property damage (with a 95% confidence interval of 0.089 – 0.230). Finally, the odds of a participant correctly judging a stimulus when visualized with the maximum-based display algorithm were 0.293 times the odds of correctly judging a stimulus displayed with the average-based algorithm.

$$
\begin{aligned}
logit\big(p(x)\big) &= \log\left(\frac{p(x)}{1-p(x)}\right) \\
&= 1.87 - 4.51x_{size} - 1.93x_{threat} - 1.23x_{display} \\
&+ 5.33x_{size}x_{threat} + 2.86x_{size}x_{display} \\
&+ 2.48x_{threat}x_{display} - 3.15x_{size}x_{threat}x_{display}
\end{aligned} \qquad (1)
$$

5. Discussion

The results show that the choice of data aggregation method did affect user errors. Furthermore, the analysis supported the hypothesis that the maximum-based display algorithm would produce a higher hit rate and false alarm rate than the average-based display algorithm. The logistic regression analysis revealed that while display condition affected task performance, participants were more accurate when judging representations of events in the low property damage category and when they had small areal coverage on the map. When evaluating the likelihood of producing a correct response at the local-level stimuli, the logistic regression analysis also showed that correctness on a national-level stimulus was a significant predictor on producing a correct response for the corresponding local-level stimulus. Likewise, congruence improved for low damage events when paired with the average-based display, whereas the maximum-based display appeared to improve congruence for high damage events. In combination, these findings suggest that choice of data aggregation technique can affect decision bias and error types.

5.1 Data Aggregation in Visual Decision Aids

The analysis of congruence by threat level and display condition suggested that the average-based display algorithm was indeed not the ideal aggregation technique. When visualizing a significant threat, the average-based display led to a divergence in judgments between the national- and local-scale stimuli. Congruent decisions increased under the average-based display condition only when the stimulus represented an event with a low property damage

level. Fidelity between the aggregated view and the local view under these conditions may be an artifact of variability in the data. When a threat is minimal, the variability in the individual data points is sometimes smaller than the variability among grid cell values for a significant threat, and thus the value produced by the average-based algorithm to represent the data aggregate at the national level more closely represents the individual members within the collection. While it is debatable whether or not a correct congruent response is more desirable than an incorrect congruent response, the results show that the display condition did affect fidelity.

With regard to incorporating data aggregation into visualization design, Elmqvist and Fekete (2010) recommend keeping the principles of visual summary and fidelity in mind. Visualizations of aggregated data ought to represent the underlying individual data points accurately and consistently. The present study's findings suggest that the average-based sampling display algorithm led to participants making significantly more congruent decisions than the maximum-based sampling display algorithm. This would indicate that the average sampling display algorithm provided a stronger representation between the aggregated, national view and the individual data points visualized in the local view. However, several visualization studies have discussed the poor ability of an average-based aggregation method to satisfy the fidelity principle. However, Elmqvist and Fekete (2010) point to a caution given by Andrienko and Andrienko (2006); they warn against using average-based aggregation methods due to the nature of averages flattening out variation. In the words of Andrienko and Andrienko (2006): "the mean weight of a fruit in a basket filled with apricots and one watermelon is… not a very useful aggregate characteristic."

Visual qualities of the stimuli shed light upon the connection between design and individual task performance. Examination of maps most often identified correctly and incorrectly showed that participants tended to correctly identify maps that represented the extremes of the stimuli (either huge swaths of floods or none at all) but had more difficulty when the maps were somewhere in the middle. The logistic regression analysis confirmed the significance of the areal

size factor within the stimuli. Misidentification of stimuli was observed in both display conditions when the stimuli sets had striking differences in visual representation between the national and local levels. For example, one stimulus contained an event that looked like a very small storm when visualized with the national-level average-based algorithm, but after zooming closer, the image actually had a very severe gradient—an indicator of flash flooding that participants were trained to seek. Participants often judged the national image to be insignificant, but changed their minds after viewing the local level. The liberal decision bias associated with the maximum-based aggregation algorithm supports the findings of Joslyn et al. (2011), who observed that providing an upper bound of forecast uncertainty led to overestimated forecasts. Joslyn et al. (2011) posited that the upper bound information provided an anchor which biased decision makers; in the current work, it is likely that the liberal bias exhibited by participants in the maximum-based condition could also be explained through this concept.

5.2 "Crying Wolf" in Weather Forecasting

Although the FLASH tools are intended for use by a population of professional forecasters and not members of the general public, aggregation methods used to create overviews influence false alarm rates and, in turn, may lead to unnecessary warnings and the "cry-wolf" effect. In the weather domain, the cry-wolf effect refers to the phenomenon wherein end-users of a weather warning fail to respond adequately after a series of false alarms, decreasing their likelihood of responding appropriately to a future true threat (LeClerc and Joslyn, 2015). In response to the concern that certain display algorithms may promote the cry-wolf effect, one must remember that selecting an appropriate response criterion is a function of signal probability and the costs of correct and incorrect responses (Wickens and Carswell, 1997). Thus, it is important to consider the cost/benefit relationship associated with response accuracy in weather forecasting. In weather forecasting, response criteria for warning on a severe threat are not only shaped by individual information processing of uncertain information, but also by governmental policy. As discussed by Doswell (2004), from a policy perspective, false alarms are often preferred over

misses, which are traditionally held in unfavorable regard. Whereas false alarms incur costs from allocating emergency response resources and may also add to a cry-wolf effect in the long run, total failure to predict a true severe weather threat can lead to significant damage and even human fatalities when protective actions are not taken. Though the present study focused on a dichotomous choice (significant versus insignificant flooding as reflected through property damage), an extension of the work could include probabilistic forecasting. If a shift in response criterion is not a viable policy option, empirical evidence is available that suggests probabilistic risk estimates attached to severe weather warnings may reduce the cry-wolf effect (LeClerc and Joslyn, 2015).

5.3 Limitations

Several factors limit the impact of this study's results. As evidenced by the possible design biases, participant judgments may have been misled by map appearances. Like many weather forecasting decision aids, FLASH is a simulated model and not a mapping of verified observations. However, to analyze the display, we showed participants FLASH maps where flooding was confirmed after the fact and then asked participants to judge whether or not they would have expected high-impact flash flooding. Therefore, participant judgments can only be as accurate as the model outputs. While we attempted to filter out FLASH outcomes of flooding events that did not appear to be accurate representations, as with any simulated model, a degree of error between the model and reality is to be expected.

In addition to modeling errors, some participants may have had difficulty gauging flash flooding severity. Participants were instructed to produce a yes or no judgment on whether or not they believed each of the displayed models could contain a severe flash flood. The definition of severe flash flooding as corresponding to greater than $500,000 worth of property damage was chosen arbitrarily in lieu of any other metric. A limitation of using property damage as a measure of severity is that it is difficult to estimate without a general knowledge about geographic features; for example, when unfamiliar with a certain region of the United States, participants

occasionally asked whether or not there were any sizeable cities impacted. Although we encouraged participants to use any background knowledge they might have of weather forecasting, we also encouraged them to make a decision ultimately based on how the FLASH stimuli appeared. In this regard, we tested the capability of the visualization to convey threat information and essentially tried to minimize the need for extensive meteorological or geographic knowledge. It is still possible that considering property damage level increased mental workload in some participants, but it is not apparent from the time-based results.

Finally, in the experimental design, a sample size issue may limit the generalizability of the logistic regression. In the original design, the variable of geographic scale was not included, but was assigned after the error rate analysis. Thus, sample sizes were uneven between the levels of the geographic scale.

6. Conclusions

The results of the present study reveal a significant difference between data aggregation display methods in terms of error types and decision bias, but not in terms of task completion time. Though the original hypothesis was that the average-based display would lead to an increase in response time over the maximum-based display, this outcome was not observed. Furthermore, the display condition did appear to affect congruence in decisions between the levels of geographic scale.

Design recommendations based on these results for future weather information displays must rely on the risk management values of the system designers. While the maximum-based display style maximized hits, it also produced many more false alarms than the average-based display. In weather forecasting, false alarms can consume valuable time that forecasters could use to analyze true threats. However, while the average-based display style produced fewer false alarms, participants were much more likely to miss an event; this could also result in critical consequences.

In the case of a flash flooding prediction system such as FLASH, the recommendation from these results would be to use the maximum-based display algorithm. Flash flooding is by nature a rapidly occurring event that can have life-threatening consequences if not predicted with enough lead-time. For such a system, having a design that promotes more hits, even at the expense of producing false alarms, would ensure that forecasters' attentions would be drawn to severe events in a timely manner. Future systems could incorporate user-configurable or automated elements. As the average-based algorithm improved congruence between the national abstraction and the local view for low damage level events, there may be utility in automating the choice of aggregation algorithm based on meteorological activity.

Whereas the present study focused on perception and comprehension, a real-time evaluation of the display methods could help to identify connections between display design and a forecaster's ability to develop SA in a dynamic manner. Additionally, future work could address limitations of the present study. While participants all had some background in meteorology and forecasting, few had specifically studied flash flood forecasting. A similar study to the present work, but run with a sample of professional flood forecasters may supplement the present study by identifying the effects of expertise on signal detection.

7. Acknowledgements

The authors would like to thank Zachary Flamig for assistance when developing the experimental stimuli and for technical support when working with the FLASH suite of products. Funding for this research was provided by NOAA/OAR/Office of Weather and Air Quality (OWAQ) under the NOAA cooperative agreement, NA11OAR4320072.

References

Adams, M.J., Tenney, Y.J., Pew, R.W., 1995. Situation Awareness and the Cognitive Management of Complex-Systems. Human Factors 37, 85-104.

Andrienko, N., Andrienko, G., 2006. Exploratory analysis of spatial and temporal data: a systematic approach. Springer Science & Business Media.

Adrienko, N., Adrienko, G., 2011. Spatial Generalization and Aggregation of Massive Movement Data. Ieee T Vis Comput Gr 17, 205-219.

Argyle, E.M., Ling, C., Gourley, J.J., 2015. Evaluation of Data Display Methods in a Flash Flood Prediction Tool, in: Yamamoto, S. (Ed.), Human Interface and the Management of Information. Information and Knowledge Design. Springer International Publishing, pp. 15-22.

Barthold, F.E., Workoff, T.E., Cosgrove, B.A., Gourley, J.J., Novak, D.R., Mahoney, K.M., 2015. Improving Flash Flood Forecasts: The HMT-WPC Flash Flood and Intense Rainfall Experiment. Bulletin of the American Meteorological Society.

Bederson, B.B., Hollan, J.D., Perlin, K., Meyer, J., Bacon, D., Furnas, G., 1996. Pad++: A Zoomable Graphical Sketchpad For Exploring Alternate Interface Physics. Journal of Visual Languages & Computing 7, 3-32.

Beven, K., Cloke, H., Pappenberger, F., Lamb, R., Hunter, N., 2015. Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface. Science China Earth Sciences 58, 25-35.

Bowden, K.A., Heinselman, P.L., Kingfield, D.M., Thomas, R.P., 2015. Impacts of Phased-Array Radar Data on Forecaster Performance during Severe Hail and Wind Events. Weather and Forecasting 30, 389-404.

Dobson, M.W., 1979. Visual information processing during cartographic communication. The Cartographic Journal 16, 14-20.

Doswell III, C.A., 2004. Weather forecasting by humans-Heuristics and decision making. Weather and Forecasting 19, 1115-1126.

Elmqvist, N., Fekete, J.D., 2010. Hierarchical aggregation for information visualization: overview, techniques, and design guidelines. Ieee T Vis Comput Gr 16, 439-454.

Endsley, M.R., 1995. Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society 37, 32-64.

Endsley, M.R., 2015. Situation Awareness Misconceptions and Misunderstandings. Journal of Cognitive Engineering and Decision Making 9, 4-32.

Fu, W.-T., Pirolli, P., 2007. SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web. Human–Computer Interaction 22, 355-412.

Gourley, J.J., Flamig, Z.L., Vergara, H., Kirstetter, P.-E., Clark III, R., Argyle, E.M., Arthur, A., Martinaitis, S., Terti, G., Erlingis, J., Hong, Y., Howard, K., 2016. The Flooded Locations And

Simulated Hydrographs (FLASH) project: improving the tools for flash flood monitoring and prediction across the United States. Bulletin of the American Meteorological Society (in press).

Hegarty, M., Smallman, H.S., Stull, A.T., 2012. Choosing and Using Geospatial Displays: Effects of Design on Performance and Metacognition. Journal of Experimental Psychology-Applied 18, 1-17.

Heinselman, P.L., LaDue, D.S., Lazrus, H., 2012. Exploring Impacts of Rapid-Scan Radar Data on NWS Warning Decisions. Weather and Forecasting 27, 1031-1044.

Hoffman, R.R., 2015. Origins of Situation Awareness: Cautionary Tales From the History of Concepts of Attention. Journal of Cognitive Engineering and Decision Making 9, 73-83.

Hoffman, R.R., Woods, D., 2005. Toward a theory of complex and cognitive systems. Intelligent Systems, IEEE 20, 76-79.

Joslyn, S., Savelli, S., Nadav-Greenberg, L., 2011. Reducing probabilistic weather forecasts to the worst-case scenario: Anchoring effects. Journal of Experimental Psychology: Applied 17, 342-353.

LeClerc, J., Joslyn, S., 2015. The Cry Wolf Effect and Weather-Related Decision Making. Risk analysis 35, 385-395.

Lehmann, A., Schumann, H., Staadt, O., Tominski, C., 2011. Physical navigation to support graph exploration on a large high-resolution display, Advances in Visual Computing. Springer, pp. 496-507.

Mays, L.W., 2010. Water Resources Engineering. John Wiley & Sons.

McNicol, D., 2005. A primer of signal detection theory. Psychology Press.

Montello, D.R., Freundschuh, S., 2005. Cognition of geographic information. A research agenda for geographic information science, 61-91.

National Climatic Data Center, 2014. Storm Events Database, in: NOAA (Ed.).

Overbye, T.J., Weber, J.D., 2000. Visualization of power system data, System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, p. 7.

Peirce, J.W., 2007. PsychoPy—psychophysics software in Python. Journal of neuroscience methods 162, 8-13.

Pirolli, P., Card, S., 1999. Information foraging. Psychological Review 106, 643-675.

Pirolli, P., Card, S., 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis, Proceedings of international conference on intelligence analysis, pp. 2-4.

Quoetone, E.M., Andra, D.L., Bunting, W.F., Jones, D.G., 2001. Impacts of technology and situation awareness on decision making: Operational observations from National Weather Service

warning forecasters during the historic May 3 1999 tornado outbreak, Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 419-423.

Ratwani, R.M., Trafton, J.G., Boehm-Davis, D.A., 2008. Thinking graphically: Connecting vision and cognition during graph comprehension. Journal of Experimental Psychology: Applied 14, 36.

Shneiderman, B., 1996. The eyes have it: A task by data type taxonomy for information visualizations, Visual Languages, 1996. Proceedings., IEEE Symposium on. IEEE, pp. 336-343.

Shrestha, A., Zhu, Y., Miller, B., 2014. Visualizing Uncertainty in Spatio-temporal data, ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA).

Smallman, H.S., Hegarty, M., 2007. Expertise, spatial ability and intuition in the use of complex visual displays, Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 200-204.

Trafton, J.G., Hoffman, R., 2007. Computer-aided visualization in meteorology. Lawrence Erlbaum.

Wickens, C.D., 2015. Situation Awareness: Its Applications Value and Its Fuzzy Dichotomies. Journal of Cognitive Engineering and Decision Making 9, 90-94.

Wickens, C.D., Carswell, C.M., 1997. Information processing. Handbook of human factors and ergonomics 2, 89-122.

Wood, E.F., Roundy, J.K., Troy, T.J., van Beek, L.P.H., Bierkens, M.F.P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P.R., Kollet, S., Lehner, B., Lettenmaier, D.P., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., Whitehead, P., 2011. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. Water Resources Research 47.

Zhang, J., Howard, K., Langston, C., Vasiloff, S., Kaney, B., Arthur, A., Cooten, S.V., Kelleher, K., Kitzmiller, D., Ding, F., Seo, D.-J., Wells, E., Dempsey, C., 2011. National Mosaic and Multi-Sensor QPE (NMQ) System: Description, Results, and Future Plans. Bulletin of the American Meteorological Society 92, 1321-1338.

Figures

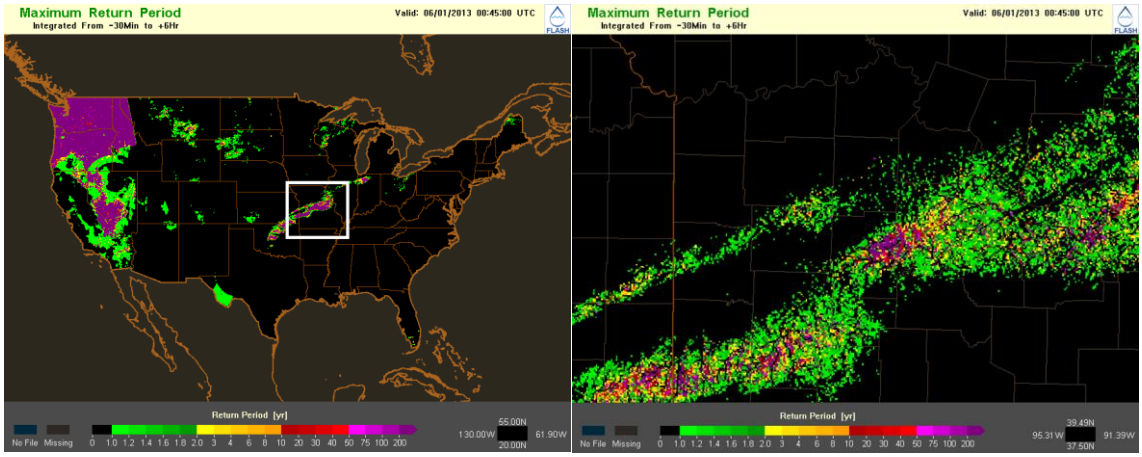

Figure 1. The national map (on left) visualized with the original maximum-based aggregation algorithm and the associated zoomed-in local view (on right)

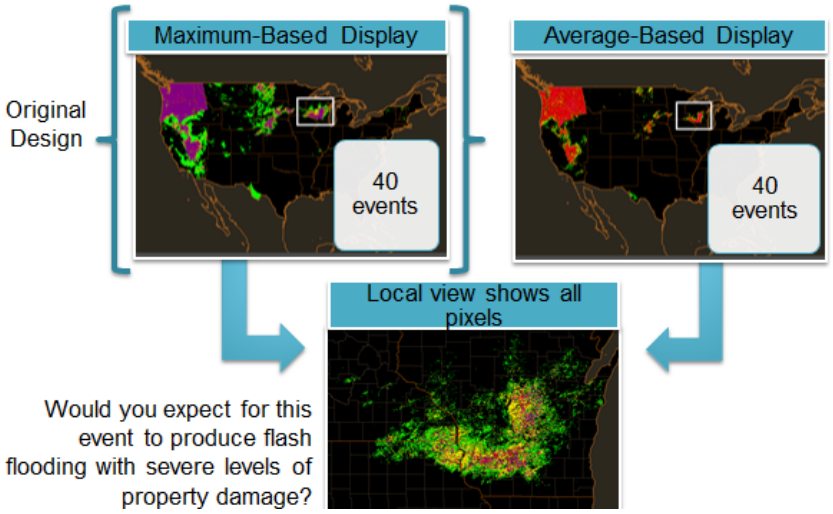

Figure 2. An example stimulus sequence, visualized with the maximum-based aggregation algorithm (on upper left) and the average-based aggregation algorithm (on upper right)

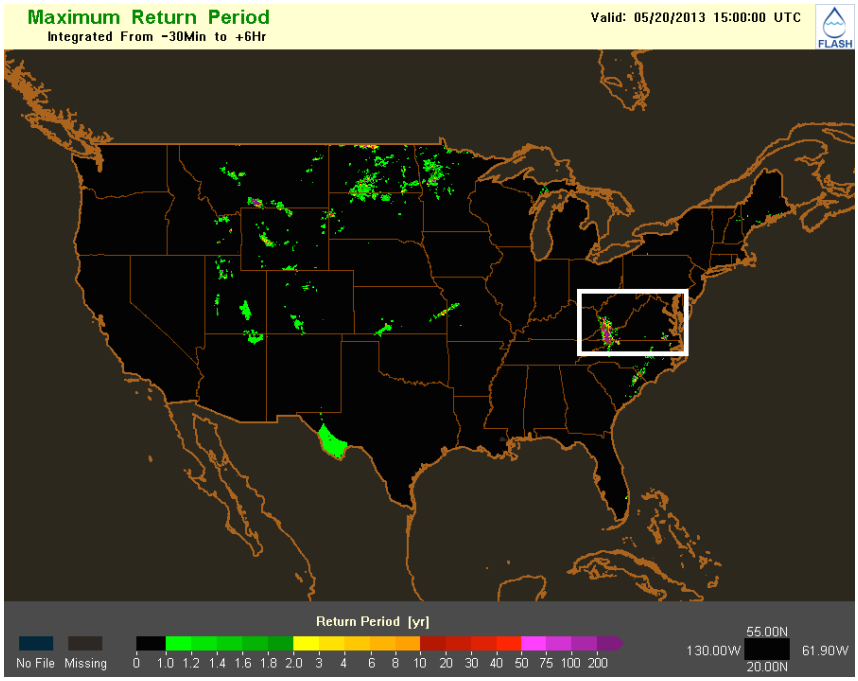

Figure 3. National-scale stimulus where size = small, threat level = insignificant (< $500,000)

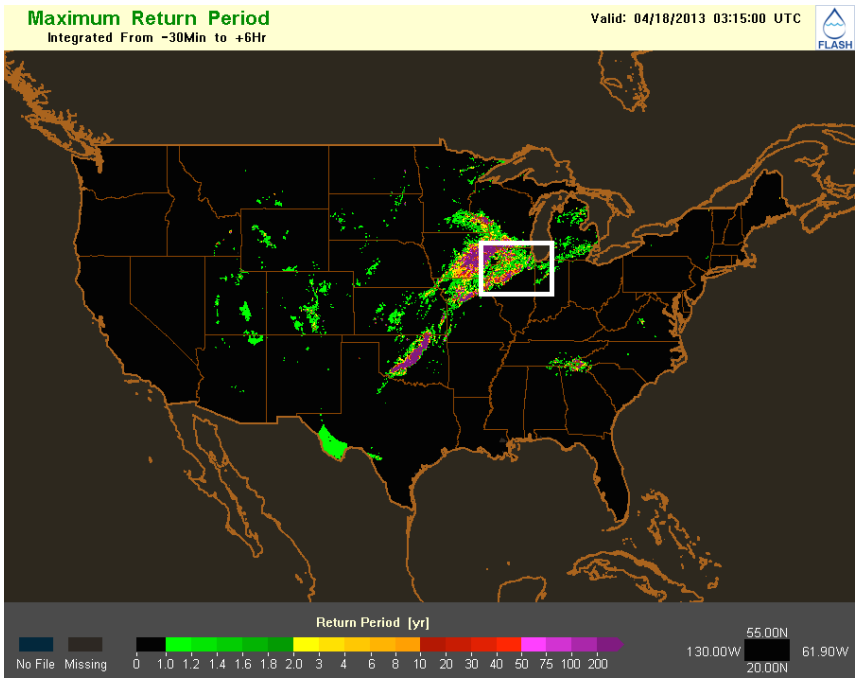

Figure 4. National-scale stimulus where size = large, threat level = significant (> $500,000)